U.S. Department of Justice
Office of Community Oriented Policing Services

COPS Publication

Community Oriented Policing Services

www.usdoj.gov/cops/

Address Based Geocoding

Final Report

Address Based Geocoding
Final Report

Police Foundation

Prepared by
Emily Powell and Michael Clifton

July 9, 1999

Final Report to the Office of Community Oriented Policing Services Cooperative Agreement # 97-CK-WX-K004

Contents

Geocoding is the process of assigning geographic coordinates to address level data for visualization and analysis in mapping software packages. In this process, specialized tools compare event data, which has been entered into a database with one or several fields used to record address information, with a map of the area of interest (reference layer). The successful matches are then used to create a new map layer with the physical location of each event represented by a point (figure 1). To ensure consistency, in packages where a new address layer is created, the new map is transformed into the same projection and attributed with the same coordinate information as the base map, including geodetic reference system information.

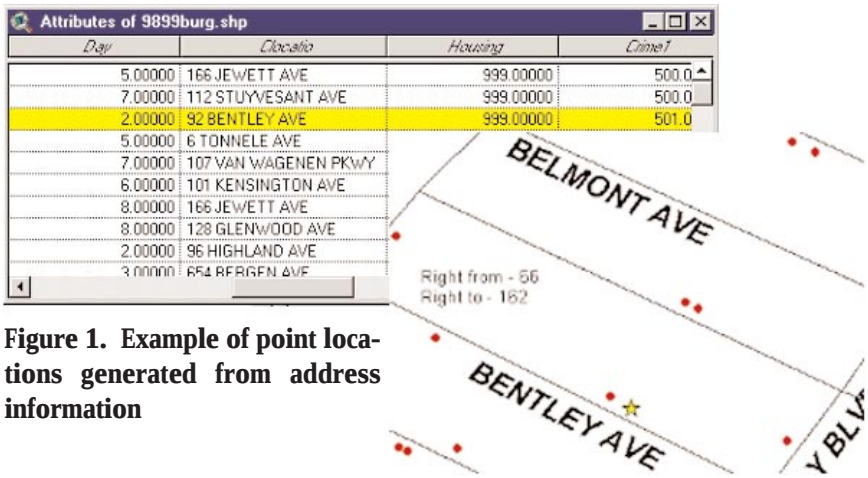

Figure 1. Example of point locations generated from address information

A variety of geographic data types may be used as a reference layer, though street files such as the Census Bureau's TIGER/line ® files are the most commonly used. In this context, street files are geographic databases containing records for a collection of interconnected street segments, defining the geography that an address matching program uses to locate a specific address. These databases generally have several fields that describe the physical address of each particular street segment in the layer, including street name, type, address numbers, prefix and suffix, and alias and often also include length information (see figure 2). The name and type are self-explanatory and are recorded in one field each. Conventionally, address numbers are recorded as 'from' and 'to' ranges for both the left and right sides of each street segment. The prefix and suffix fields are used to store information such as quadrant and other abbreviated identifiers; for example, 155 W Hollywood Dr. would include both a prefix (W for west) and a suffix (Dr. for Drive) as well as the street name and number.

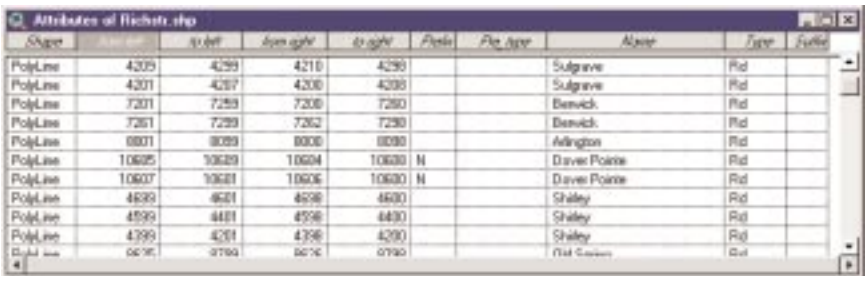

Figure 2: Sample Street File Attribute Table

A similar geocoding strategy uses polygons as opposed to previously discussed linear reference layers. This method is similar to address geocoding in that a value from the geocoded data must match that of an existing layer. For example, calls for service data often has a grid, beat, or district field in addition to an address field. The incidents can then be geocoded against a corresponding grid, beat, or district layer (provided there is one). Other types of polygon data are zip codes, census blocks, census tracts; data can also be geocoded by city, county, state, and country. This type of geocoding is often used to create thematic maps. However, when data are geocoded by polygon specific address information will not be available for mapping. This is why many people geocode the address level and aggregate the data posthumously. Geocoding data at the polygon level is particularly useful when a data set is very large or when address information is not as reliable as the grid, beat, etc. information.

Parcel files are also polygon files. They contain a shape for every parcel of land that have a parcel number and an address. When geocoding by parcel address (called situs address), the address must match to the exact number and street. There are no ranges by which to match. This form of geocoding is more accurate than geocoding by centerline maps because the exact address is located; however, there is a higher threshold for matching, and if maps or data aren't accurate, the number of successfully geocoded addresses is lower.

A different, but also common geocoding strategy uses polygons as opposed to the previously discussed linear reference layers. This method creates a centroid point attributed with the corresponding address within each polygon, and then matches the polygon and event addresses. This approach is commonly used for aggregating point information to polygon areas such as census block groups or postal code zones. Geocoding by any user-defined political or geographic region may be possible. Parcel files, comprised of polygon boundaries for buildings in an area, are often used in place of street lines; because like intersection geocoding, it eliminates the inaccuracy introduced by mathematically deriving location and also avoids the problem of imprecise street files.

An assessment of which reference layer would be most effectively employed must consider the needs of the individual end user. Considerations of accuracy, cost of product, maintenance costs, utility, current reporting procedures, and comparability with other existing reference layers must be dealt with on a case by case basis. In the case of law enforcement, some departments are primarily concerned with more densely populated areas with a large amount of development and, therefore, change. For these departments, it might be wise to develop a partnership with other local agencies which would have use for the same reference layers, therefore sharing the cost of the frequent updates needed to maintain accurate geocoding. The need for local agencies to update street files stems from urban growth and the building or changing streets as well as because national updates are very slow and less accurate.

Because the Census Bureau's mission to count and profile the Nation's people and institutions does not require very high levels of positional accuracy in its geographic products, its files and maps are designed to show only the relative positions of elements. Some level of correction is generally necessary when using these products. Many private firms and local governments spend time and resources updating and correcting these files for particular areas for sale or for their own use, creating products known as "modified TIGER ®".

One alternative product which is currently being studied and implemented is the GPS (Global Positioning System) centerline file. A GPS centerline is produced by driving the existing roadways and recording coordinate points at pre-determined intervals using a GPS receiver. Because of recent advances in differential GPS technology and the inherent distortion in remotely sensed images and errors in the digitizing process, these files can easily rival the accuracy of traditional centerline files. An added advantage to this type of centerline file is that when new development occurs it can be added to the pre-existing centerline file. Currently, several companies are creating custom GPS centerline files at the regional and county level.

As previously mentioned, another common layer used as a reference for geocoding is the parcel file. A parcel file is a polygon layer, often developed originally by or for local governments, used to keep track of lots and subdivisions for planning and tax purposes (see figure 3). These are most often created from planimetric layers originating from aerial photography or by digitizing planning documents such as site plans and blue prints using CAD (Computer Aided Design) software packages. Parcel files provide a good base layer to examine relationships between land-use, zoning, demographic data, and economic data. While parcel files have the ability to be highly accurate, like street files, this accuracy is largely dependent on how often and how well they are maintained. Because of the nature of real estate and other planning mechanisms, keeping parcel files updated and accurate can be a full time job for several technicians depending on the area.

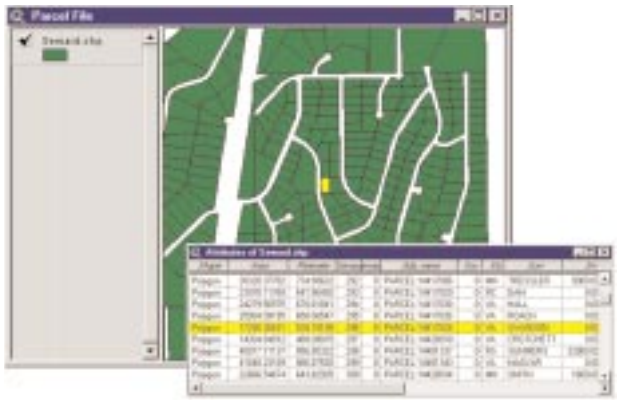

Figure 3: Sample Parcel File and Associated Attribute Table

Successful geocoding relies on the quality of not only the reference layer, but also the data to be mapped. For any crime mapping application to be effective, basic reporting procedures need to be uniform and consistent. While most high-end geocoding software packages incorporate complicated algorithms making them somewhat forgiving, a well established reporting procedure is often the best way to ensure a high rate of accuracy. In the past when technology was not as prevalent as it is today, various agencies and even sections or individuals within agencies would use their own formats for addressing and naming places. Currently many departments whether already have in place or plan to acquire an automated computer aided dispatch (CAD) system and a record management system (RMS) which have the compatibility to store digital information as well as standardized tables of addresses and location names. However, clear and consistent rules must be established in order to ensure that data are reliable and valid. In addition, because individuals are the data entry mechanism for the data, there is still the possibility of human error and inconsistencies, which is why training should coincide the technology.

Incorrect information or careless data entry can make geocoding impossible. Address errors are often the source of much of the woes of geocoding and can be categorized into five common types: incorrect street numbers, street name errors, direction errors, and incorrect intersections (in the case of intersection geocoding). Street number errors can be the result of entry mistakes and are frequently identified as range errors; the address identified is either less than or greater than the address range of the corresponding records in the source file. Street name errors can occur for a variety of reasons, and are a result of the address record not matching the street name in the source database. These errors are among the most common, with reasons including misspellings, inconsistent street types, or lack of compliance with City addressing standards. Direction errors occur when the direction code of the address record does not match the direction code in the source database, either because they are incorrect or missing. Incorrect intersections can result from data entry problems. For instance- Baltimore & Eutaw will not be recognized by many packages if entered as Baltimore/Eutaw or Eutaw & Baltimore.

Before geocoding, there are many possible procedures for correcting address errors. Manual correction may be possible for small projects, or in instances where multiple records contain the same error. Data sets can often be corrected more efficiently when imported into a database or spreadsheet program (for example, Excel or Access) which has a search and replace utility. Specialized software has been customized to automatically clean up data sets when manual repair is not practical. These specialized programs, referred to as "scrubbers", can be useful for large projects, and in instances where data is regularly imported. For reoccurring problems, many GIS packages offer the option of creating an alias table, which allows for multiple names for a single entity.

Logically, the other set of geocoding problems results from errors in the reference layer. These are generally manifested in the form of missing address ranges or even complete street segments. Unmodified TIGER® files often lacking address ranges for streets because of recent development, and oversights in the original creation process. In urban areas, the percent of street segments that contain address ranges may be as high as 90+ percent. Given the odds, the success rate for address matching a database of points will vary dramatically depending upon the study area. Because street files are generally manually attributed, they are also prone to data entry errors. The combination of these problems can be difficult to solve, and may require the purchase of more expensive commercially produced enhanced files. Many agencies have opted either to purchase street files that have been corrected and are bundled with contracts for regular updates or to produce and maintain their own set. In either case, interagency cooperation can prove invaluable as a means for making the geocoding process feasible.

The ArcView address geocoding process requires that a reference theme, a geographic street layer (in shape file format) and a table (in DBF, INFO or delimited text format) containing the addresses to be located are loaded into an ArcView project.

1. Before the geocoding process begins, it is advisable to check that the theme properties for the street theme are set correctly. This can be found by selecting "properties" from the "Theme" menu while the street theme is selected in the view document.

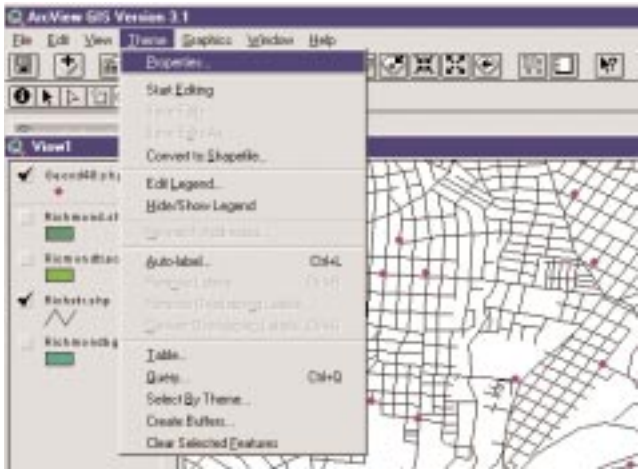

Figure A.1.

2. The Geocoding portion of the "Theme Properties" dialog window allows the user to select which fields in the street file's table contain the information necessary for geocoding. The option to choose to use the suffix for a street name is also provided, and must be considered for optimal address matching. Drop down menus , each of which contain a list of fields in the street file, allow easy reassignment of necessary information.

Figure A.2.

3. The geocoding process is launched when the user selects "Geocode Addresses" from the "View" menu.

4. In the Geocoding dialog box, the street file must be selected in the "Reference Theme" section, and the table containing the addresses to be used should be selected in the "Address Table" section. The field containing the address information must be then selected from a scrollable list of fields in the chosen table. A default name is created for the geocoded point theme that will be generated. The user may change this name in the "Geocoded Theme" box.

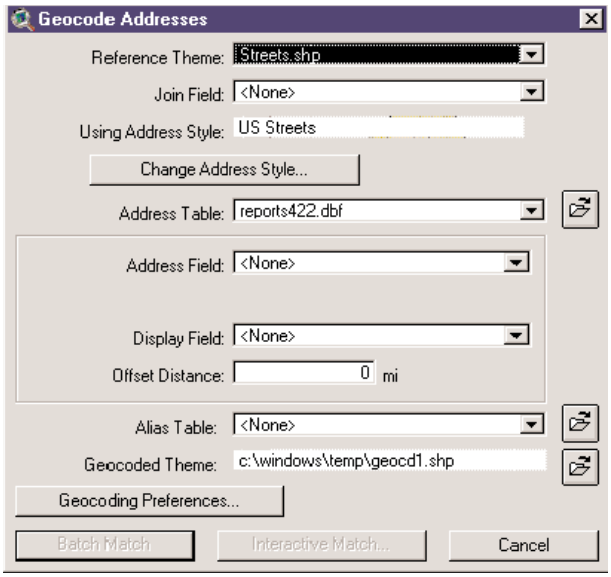

Figure A.3.

5. The "Geocoding Preferences" accessible from the Geocoding Addresses dialog box allow the user to define matching sensitivities considered when candidates are being chosen during the matching process. Preferences may also be changed later in the geocoding process. ArcView allows the user to choose to interactively match addresses or to elect to have the program make a "batch match" of all addresses that it considers recognizable. In the case of batch match it is more important that the sensitivities are set high to avoid erroneous matches. To begin either process, select one of the buttons at the bottom of the "Geocoding Addresses" window.

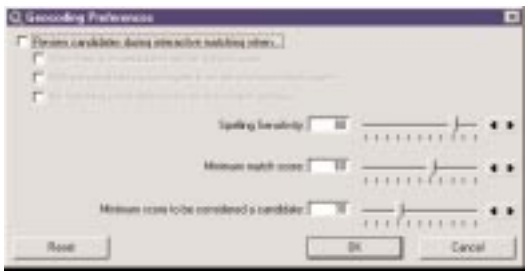

Figure A.4.

6. The Interactive matching process opens up a "Geocoding Editor" window which displays addresses in order, one at a time, and any matches that fit into the criteria selected in the "Geocoding Preferences" (which may be altered at this point). To match records, the user selects the candidate that matches the most correctly, and clicks the "Match" button. The editor then automatically advances to the next record.

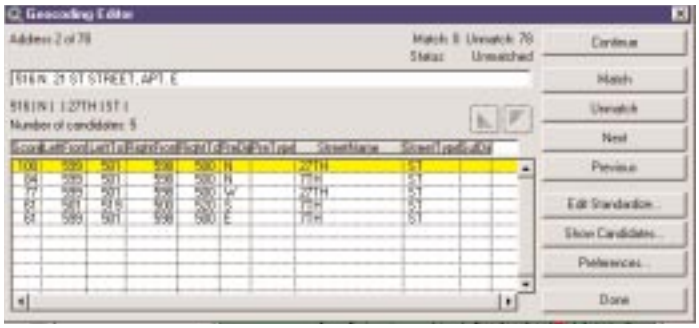

Figure A.5.

7. In the event that there are no matches automatically displayed, the "Edit Standardize" option allows changes to be made to the address format and spelling. There are several likely causes of data mismatch that can be interactively corrected at this point. One common error is that a street name has a component thatcan be confused with another of the categories that the geocoding editor automatically breaks the complete name into. This often happens with names like "River Run ct." where the "Run" might be mistaken for the street type instead of part of the name. Another common problem is when a street's name contains a direction. When the editor encounters a street named something like "N Ridge Rd." or "East Lombard St." it assumes that the "N" or "East" is the direction of the street, making a match without standardizing is unlikely.

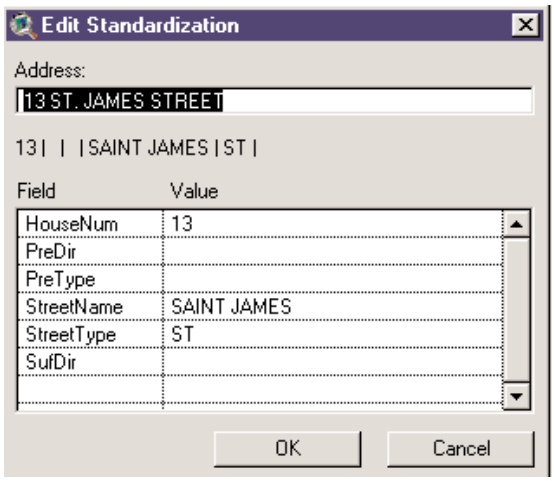

Figure A.6.

8. When geocoding is complete, ArcView displays a summary dialog. The user can choose to re-match at this point or later can choose "Re-match Addresses" from the "Theme" menu. The geocoded table is also changed to reflect the results of the geocoding process. Five fields are automatically added to the original data table containing information about the address used to make the match: the address actually used to mathc, used the zone used (if this is a possibility), the status of the geocoding (M if matched, U if not), the score assigned to this match, and the side of the street that the address falls on.

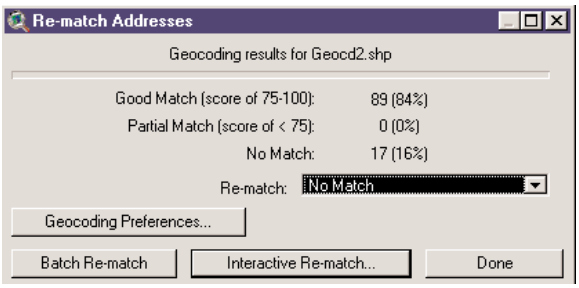

Figure A.7.

To geocode using the MapInfo GIS package, at least one table containing geographic information to be mapped (in one of a variety of standard formats), and one table containing mappable geographic source information (for instance a street file) must be open in a MapInfo workspace.

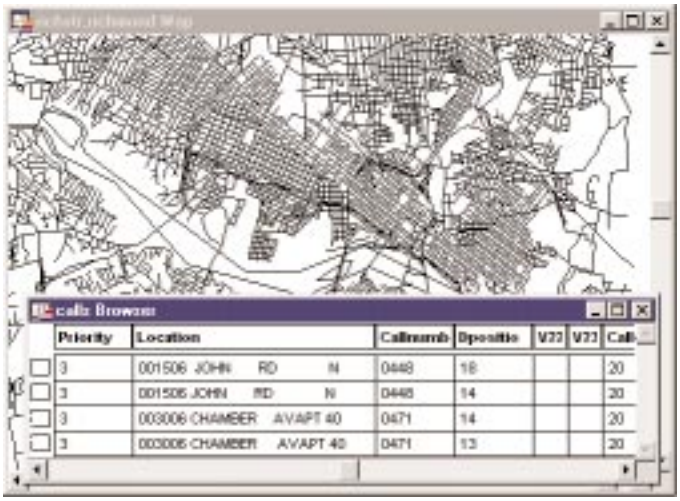

Figure B.1.

1. The tables used in the geocoding process must be configured correctly before geocoding using MapInfo. Open the "Table Structure" option in the "Maintenance" portion of the "Table" menu. From the "View/Modify Table Structure" select the appropriate table.

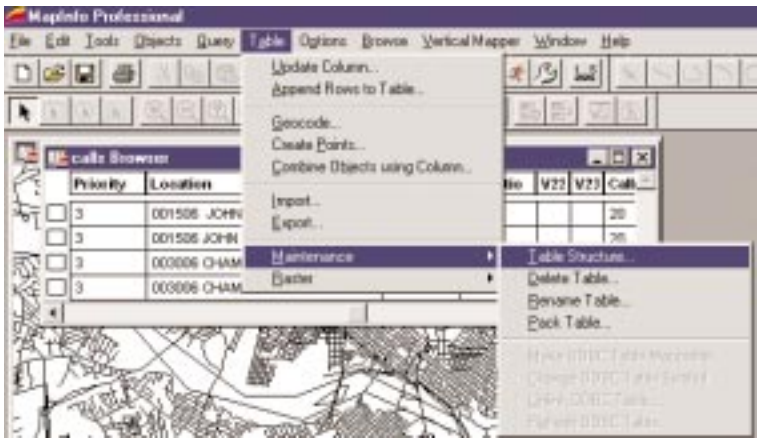

Figure B.2.

The field containing the street name information in the reference layer must be indexed by checking the box to the right of the field label in the "Modify Table Structure" dialog, and must appear before the from and to address fields.

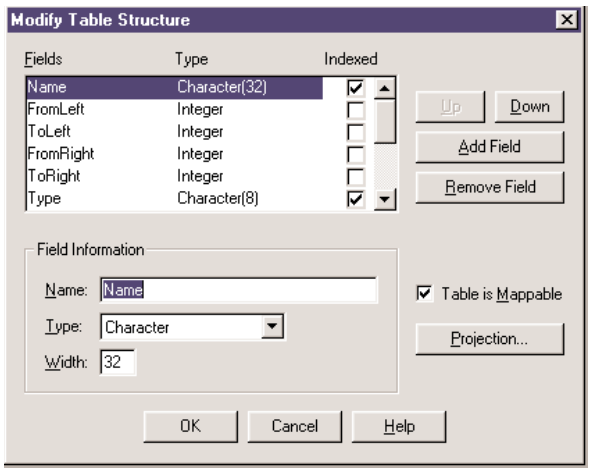

Figure B.3.

The structure of the table that is being geocoded should also be altered using the "Modify Table Structure" dialog procedure. A numeric column to receive the geocoding status that MapInfo automatically adds should be added to the table prior to geocoding.

2. To begin the geocoding process, select "Geocode" from the "Table" menu.

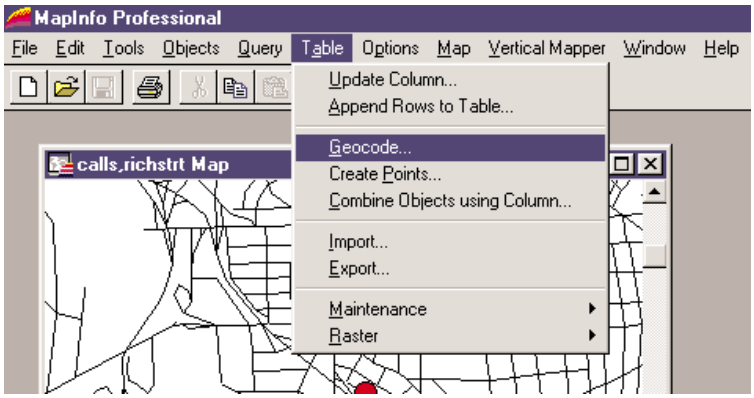

Figure B.4

3. In the Geocode dialog box, the table containing the data to be geocoded should be chosen from the drop-down menu in the "Geocode Table" section, and the "using Column" box should contain the name of the field that contains the address information. The "Boundary Column" is an optional field that can be used to specify an additional field with information to distinguish between like-named segments in different areas, for instance, the individual municipalities for each street. The Search Table field should display the name of the reference table. The "for Objects in Column" field should contain the previously indexed name field. The optional section of the form can be used to specify more boundary information contained in a separate table. Before geocoding, the symbol that

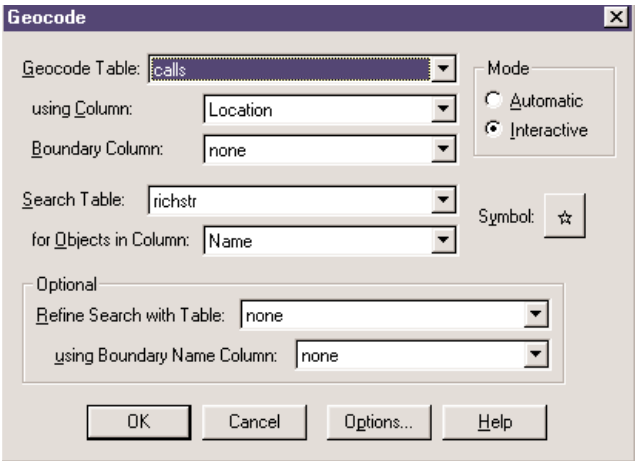

Figure B.5.

will be used to display the new points can be chosen. The decision to use the "Automatic" or "Manual" modes, where the software automatically matches addresses or the user chooses the appropriate match, should be made based on the size of the record set and the ability of the software to make automatic matches based on available data. The MapInfo documentation recommends that it may be advisable to begin with automatic mode, and after reviewing the results, manually geocode instances that do not produce an automatic match.

4. MapInfo presents a simple interface for interactive geocoding. In the interactive mode, it is possible to edit the address text in the "Name" field, so that the geocoding process searches for an edited version of the address. This can be useful in instances where there are obvious flaws in the address format. The "Up" and "Down" buttons are used to scroll through the available records, and the "Ok" and "Ignore" buttons offer the choice to select a match or to continue without matching the selected record.

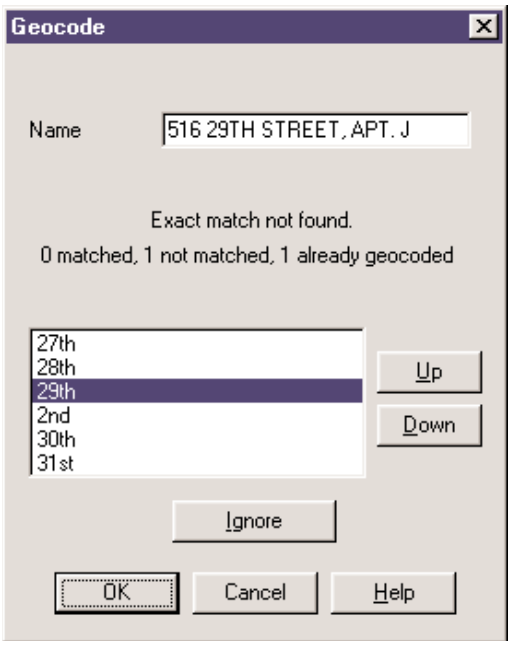

Figure B.6.

5. After geocoding by either method is complete, either in interactive or automatic mode, a dialog displays the results of the procedure. If the records were not all successfully matched, the user may choose to change parameters or edit the files to correct errors before running the geocoding procedure again.

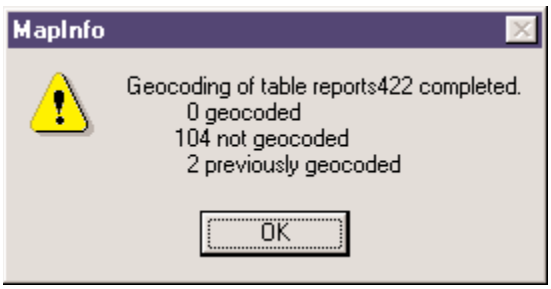

Figure B.7.

For More Information:

U.S. Department of Justice
Office of Community Oriented Policing Services
1100 Vermont Avenue, NW
Washington, D.C. 20530

To obtain details on COPS programs, call the
U.S. Department of Justice Response Center at 1.800.421.6770.

Visit the COPS internet web site:
www.usdoj.gov/cops

Created: September 29, 1999
e09990008